Acceptance Test for the Intel Paragon XP/S-15

Bernard Traversat[1] and David McNab[1]

Report RND-94-004 February 1994

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

ARC 275 (Rev Feb 81)

Acceptance Test for the Intel Paragon XP/S-15

Bernard Traversat[1] and David McNab[1]

Report RND-94-004 February 1994

travers@nas.nasa.gov
NAS Systems Development Branch
NAS Systems Division
NASA Ames Research Center
Mail Stop 258-6
Moffett Field, CA 94035-1000

Abstract

On February 17th 1993, the Numerical Aerodynamic Simulation (NAS) facility, located at NASA Ames Research Center, installed a 224 node Intel Paragon system. After its installation, the Paragon was unable to complete any significant tasks without crashing. A simple "Hello World" program, in which each node printed the words "Hello World", froze the system when run on more than 16 nodes. Uptime was less than fifteen percent with approximately ten reboots per day. No acceptance test was run on this date.

On July 22nd 1993, after functionality and stability improved an acceptance test was run. The motivation for this test was to have the Paragon demonstrate a minimum level of stability and performance before committing further development resources. The acceptance test consisted of compiling and running simultaneously two copies of the NAS Parallel Benchmarks for a period of twelve hours. The test was considered to pass if all the NAS Parallel Benchmarks ran faster than reported results for the iPSC/860, and if less than two crashes and greater than ninety percent uptime, per twelve-hour period, were achieved for seven consecutive days. The acceptance criteria were deemed to be the minimum needed to open the Paragon to a small set of users, and justify an attempt to fix the remaining problems. On August 2nd 1993, due to numerous system software failures the test was postponed.

On November 16th 1993, after OS release R1.1 was installed and newer versions of the NAS Parallel benchmarks were available, the acceptance test was resumed. Two weeks later, after the test requirements were met the test passed. Although the acceptance test was a success, low stability and functionality still make the Paragon unable to support a real workload.

---

1. Computer Sciences Corporation, NASA Contract NAS 2-12961, Moffett Field, CA 94035-1000

1.0 Introduction

A goal of the Numerical Aerodynamic Simulation (NAS) facility is to have highly parallel computer systems support a workload similar to the one currently run on its conventional vector supercomputer systems (i.e. Cray C90). This workload is composed of a wide range of development and production activities involving large scale Computational Fluid Dynamics (CFD) computation. Interactive and batch jobs are mixed, and a system is commonly shared between a large number of simultaneous users (~100). On February 17th 1993, a 224 node Intel Paragon XP/S-15 was installed at NAS, to complement the 128 node iPSC/860 and 128 node CM-5 testbed parallel systems. Like the other testbed parallel systems, the Paragon was not required to pass an acceptance test. However, after installation the Paragon was found to be in a very immature state [5]. Uptime was less than fifteen percent with approximately ten reboots per day. Serious hardware and software problems, such as node board failures, virtual memory thrashing and process management corruptions made the system unusable. The system was unable to complete any significant tasks without crashing or hanging. A simple "Hello World" program, in which each node printed the words "Hello World", froze the system when run on more than 16 nodes. The Embarrassingly Parallel (EP) benchmark was the only NAS Parallel Benchmark that would even run sporadically on the system. In order to commit further development and support resources, a "*minimal*" acceptance test was run. The motivation for this test, was to have the Paragon demonstrate a minimum level of functionality, stability and performance under a very low workload (i.e. 2 users). The acceptance criteria were deemed to be the minimum required to open the system to a small set of users, and justify an attempt to fix the remaining bugs. If this test could not pass, the system would be considered as unusable and support would be terminated.

This report describes the attempts made by NAS and Intel Supercomputing Systems Division (SSD) to have the Paragon pass this acceptance test. A brief description of the Paragon configuration, installed at NAS, is given as well as the test procedure and acceptance criteria. The log summaries of the two attempts (July 22th and November 16th) are presented, and the report concludes with overall impressions on the acceptance test outcome.

2.0 The Intel Paragon XP/S-15

The Paragon XP/S-15 [1] is a distributed-memory multiprocessor system using a two dimensional mesh interconnection network composed of 224 nodes. Each node consists of two Intel i860 XP microprocessors (one used for computation and one intended for communication) with 16-32 MBytes of local memory. The communication coprocessor could not be used during the test period. The i860 XP runs at 50MHz with a 75 MFLOPS (double precision) peak performance. The mesh routing hardware is capable of delivering a node-to-node peak bandwidth of 175 Mbytes/s (full duplex).

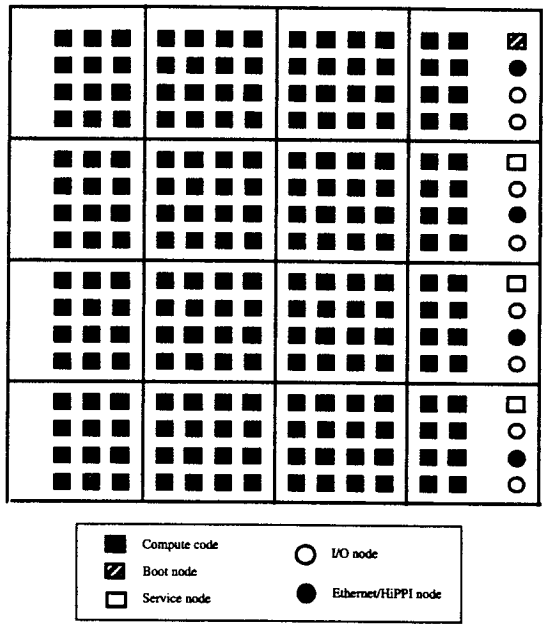

Figure 1. Paragon XP/S-15 Configuration

The Paragon Operating System (Paragon OS) is based on the Open Software Foundation Advanced Development operating system (OSF/1 AD) [2,3]. OSF/1 AD is a distributed-memory operating system based on the Mach microkernel from Carnegie Mellon University and the OSF/1 Unix implementation. The Paragon XP/S-15 configuration, installed at NAS, has two hundred and eight compute nodes (16 Mbytes of memory each), four service nodes, three Ethernet and HiPPI nodes (16 Mbytes). Eight additional I/O nodes are attached to RAID disk drives for a capacity of 38 Gbytes of usable secondary disk space. On the first attempt to run the acceptance test, on July 1993, only one service node (boot node) had 32 MBytes. During the second attempt, on November 1993, all four service

nodes were upgraded to 32 MBytes. The service nodes serve as "front-ends" to the system's compute nodes, and provide traditional Unix interactive facilities such as editing, compiling and program execution.

3.0 Acceptance Test Procedure

The acceptance test was designed to demonstrate a minimal level of functionality, stability and performance. The test consisted of compiling and running simultaneously, two copies of the NAS Parallel Benchmarks (NPB) [4] for a period of twelve hours. The NAS Parallel Benchmarks are a set of 5 kernels and 3 application benchmarks that represent computational parts of important NAS application codes. The *class A* size of the benchmark specifications was used, as the larger *class B* size was not available at the time of this test. The NPB implementations were provided by Intel SSD.

The test was run for a twelve-hour period from 9:00 PM to 9:00 AM. The twelve-hour period was considered sufficient to evaluate system reliability, while leaving enough hours for system development work. The test ran in dedicated mode with no interactive users. The Paragon was rebooted before each test to start with a clean system. Memory leaks existing in the OS were known to degrade system functionality after it had been running for several hours. Two simultaneously running copies of the script file, given in Figure 2, were run. Two running codes was the minimal number needed to test the node allocator and process scheduler under a multi-user workload. The two scripts ran asynchronously and on different directories to avoid overwriting files.

The eight NPB's ran for different node configurations (see Figure 2). All the benchmarks, other than EP, had limitations on the number of nodes on which they could run. They also had to be compiled for each node configuration. The need to recompile the benchmarks before each run was found to be a good feature, as this effectively stressed the service partition and disk I/O which would have been otherwise idle. Most benchmarks were run on 64 and 128 node partitions. Only the EP and APPLU benchmarks were able to run on the entire 208 node compute partition. Due to resource scheduling conflicts, if one job was running on 128 nodes, the other job could not run on 128 nodes since only 208 nodes were available. Depending on the OS release, the second job was either blocked until the requested resources were released (T10) or the job would fail and the script would continue to the next procedure step (R1.1). The `-sz` scheduling option was used to specify the number of nodes requested to run an application. This option safely prevented the execution of applications on overlapping nodes.

```
/*
EP Benchmark (Size 2^28)
*/
make EP
run EP on 32 nodes
run EP on 64 nodes
run EP on 128 nodes
run EP on 208 nodes
/*
APPBT Benchmark (Size 64^3)
*/
make APPBT for 64 nodes
run APPBT on 64 nodes
Make APPBT for 128 nodes
run APPBT on 128 nodes
/*
APPSP benchmark (Size 64^3)
*/
make APPSP for 64 nodes
run APPSP on 64 nodes
make APPSP for 128 nodes
run APPSP on 128 nodes
/*
APPLU benchmark (Size 64^3)
*/
make APPLU for 64 nodes
run APPLU on 64 nodes
make APPLU for 128 nodes
run APPLU on 128 nodes
make APPLU on 208 nodes
run APPLU on 208 nodes
/*
MG benchmark (Size 256^3)
*/
make MG
run MG on 128 nodes
/*
FFT Benchmarks (Size 256^2x128)
*/
make FFT for 64 nodes
run FFT on 64 nodes
make FFT for 128 nodes
run FFT on 128 nodes
/*
IS benchmark (Size 2^23)
*/
make IS
run IS on 32 nodes
run IS on 64 nodes
run IS on 128 nodes
/*
CG benchmark (Size 2.0x10^6)
*/
make CG
run CG on 128 nodes
```

Figure 2. NPB Test Script

The acceptance test would pass, if the eight NPB's implementations ran faster than reported results for the iPSC/860, and if less than two crashes and greater than ninety percent uptime per twelve-hour period were achieved for seven consecutive days. The performance criteria between the Paragon and the iPSC/860 was selected to verify no performance degradation as the Paragon was expected to eventually replace the iPSC/860. The stability criteria was the minimum thought to insure that the system stayed up long enough so users could accomplish work. A monitoring procedure checked the output of the jobs every ten minutes for possible deadlock or failures.

4.0 July 22nd 1993 Attempt

On July 22nd 1993, after the performance criteria was met for the first time, an attempt was made to run the acceptance procedure. At that time, the OS release T10 and the new firmware (Fab7-11) were installed on the system. The Paragon functionality had improved from merely running one of the NAS Parallel Benchmarks to running all of them between 1 and 2 times faster than on the iPSC/860 (see Table 1). Uptime was around eighty percent with an average of three to four reboots per day [5].

TABLE 1. NAS Parallel Benchmarks (T10)

Benchmarks	Problem Size (Class A)	128 Nodes T10 (Secs)	128 Nodes Ratio to iPSC/860[1]
EP	2^{28}	19.72	1.30
MG	256^3	6.52	1.32
FFT	256^2x128	6.36	1.52
CG	$2.0x10^6$	6.75	1.27
IS	2^{23}	13.50	1.01
APPSP	64^3	281.27	1.59
APPBT	64^3	266.58	1.56
APPLU	64^3	442.00	1.01

[1] iPSC/860 Timings from the "NAS Parallel Benchmarks Results", D. Bailey et al.,NAS Technical Report RNR-92-002.

TABLE 2. July 22nd 1993 Attempt Run

Date	Number of Reboots	Uptime (%)	Test Pass/Fail
7/22/93	2	85.2	*Fail*
7/23/93	4	71.5	*Fail*
7/24/93	2	83.8	*Fail*
7/25/93	1	93.6	Pass
7/26/93	3	65.4	*Fail*
7/27/93	2	87.7	*Fail*
7/28/93	3	81.6	*Fail*
7/29/93	9	49.3	*Fail*
7/30/93	9	57.7	*Fail*
7/31/93	2	86.3	*Fail*
8/01/93	3	66.7	*Fail*

On August 2nd after consistent failures, the test was postponed until OS stability improved. Two or more system software crashes were regularly observed per twelve-hour run. Outstanding bugs, like virtual memory leaks, OS inter-process communication deadlocks, and process management corruptions were limiting critical OS functionalities such as multi-user support, virtual memory paging, and system reliability. During the test period no hardware failures were experienced, all the reboots were due to system software failures.

5.0 November 16th Attempt

On November 16th 1993, after the second official OS release (R1.1) was installed, all four service nodes were upgraded to 32 MBytes, and newer versions of the benchmarks were available, the acceptance test was resumed. The NPB's ran between one and three times faster than on the iPSC/860 (see Table 3). Uptime was around eighty-five percent with an average of three reboots per day. Multi-user support and process management improved, but the system was still having OS problems.

TABLE 3. NAS Parallel Benchmarks (R1.1)

Benchmarks	Problem Size (Class A)	128 Nodes T10 (Secs)	128 Nodes Ratio to iPSC/860[1]
EP	2^{28}	17.01	1.51
MG	256^3	5.66	1.52
FFT	256^2x128	6.26	1.55
CG	2.0x10^6	6.31	1.36
IS	2^{23}	13.50	1.01
APPSP	64^3	208.13	2.16
APPBT	64^3	147.60	2.81
APPLU	64^3	383.75	1.15

[1] iPSC/860 Timings from the "NAS Parallel Benchmarks Results", D. Bailey et al., NAS Technical Report RNR-92-002.

TABLE 4. November 16th Attempt Run

Date	Number of Reboots	Uptime (%)	Test Pass/Fail
11/16/93	4	85.4	***Fail***
11/17/93	4	82.1	***Fail***
11/18/93$^{\alpha}$			
11/19/93	1	97.8	**Pass**
11/20/93	3		***Fail***
11/21/93	0	100	**Pass**
11/22/93	0	100	**Pass**
11/23/93	1	99.4	**Pass**
11/24/93	2	92.5	**Pass**
11/25/93	1	98.3	**Pass**
11/26/93$^{\alpha}$			
11/27/93	2	98.6	**Pass**
11/28/93$^{\alpha}$			
11/29/93	1	97.0	**Pass**

[$^{\alpha}$: Test was not run correctly due to operational failures unrelated to Paragon stability.]

On November 30th 1993, after seven consecutive successes, the acceptance test was passed. The failures that occurred before November 18th, were due to bad node hardware. After those nodes were replaced, the remaining crashes were due to system software failures. The failures were traced to bugs still remaining in the OS, such as virtual memory leaks and OS inter-process communication deadlocks.

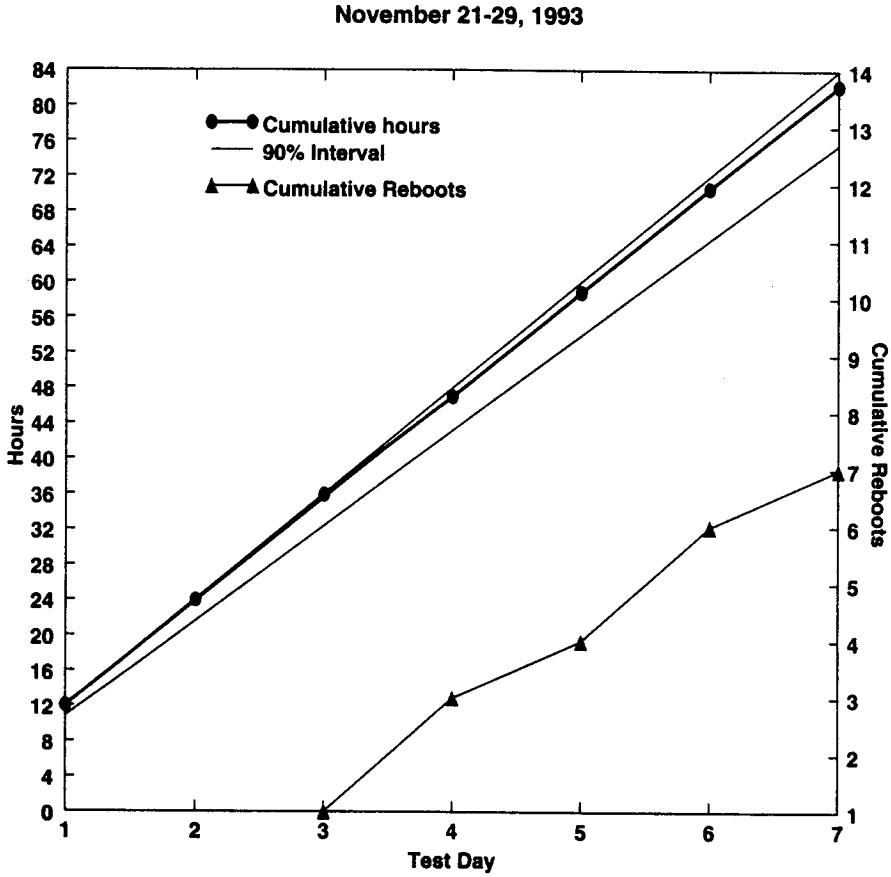

Figure 3. Cumulative Uptime and Reboots

6.0 Summary

On February 17th 1993, a 224 node Paragon XP/S-15 was installed at NAS. Upon installation, the system was found unable to complete any significant tasks. Uptime was less than fifteen percent with approximately ten reboots per day. No acceptance test was run at this date.

Five months later, after all the NPB's were able to ran faster than on the iPSC/860, an attempt was made to run an acceptance test. The test consisted of compiling and running simultaneously two copies of the NAS Parallel Benchmarks, for a twelve-hours period. On August 2nd 1993, due to consistent system software failures, the test was postponed. More than two crashes per twelve-hour period were observed.

On November 16th 1993, after OS release R1.1 was installed, all four service nodes were upgraded to 32 MBytes, and newer versions of the NPB's were available the acceptance test was resumed. On November 30th 1993, after the system met the seven-day requirements the test passed.

Although the acceptance test succeed, the two attempts showed that the Paragon had problems running this minimal workload due to serious system software limitations. It took almost nine months for the Paragon to pass this minimal test. At the end of the test, the system still crashed consistently at least once per night under very little workload. The low observed stability makes the Paragon unable to support a real user workload. The improvement in system stability between the first and second test attempts appeared to be due to the upgrade of all four service nodes to 32 MBytes rather than an improvement in system software reliability. The same OS bugs have been consistently observed during both attempts. Until these critical OS bugs (virtual memory leaks, process management corruptions, and OS inter-process communication failures) are fixed, limited improvements in stability should be expected.

7.0 Acknowledgments

The authors wish to thank the NAS Systems Control Staff for monitoring the test during night hours, and Ed Kushner and Thanh Phung from Intel SSD for providing the NPB's implementations.

8.0 References

[1] "Paragon OSF/1: User's Guide", *Intel Supercomputer Systems Division*, April 1993, Order Number: 312489-001.

[2] "An OSF/1 Unix for Massively Parallel Multicomputers", R. Zajcew et al., *in Proceedings of the 1993 Winter USENIX Conference*, January 1993, pp. 37-55.

[3] "OSF Mach: Kernel Principles", K. Loepere, Open Software Foundation and Carnegie Mellon University, February 1993.

[4] "The NAS Parallel Benchmarks", D. Bailey et al., NASA Technical Memorandum 103863, *NASA Ames Research Center*, July 1993.

[5] "Evaluation Metrics for the Paragon XP/S-15", B. Traversat et al., NAS Report RND-93-017, *NASA Ames Research Center*, December 1993.

RND TECHNICAL REPORT

Title: Acceptance Test for the Intel Paragon XP/S-15

Author(s): Bernard Traversat and David McNab

Reviewers:
"I have carefully and thoroughly reviewed this technical report. I have worked with the author(s) to ensure clarity of presentation and technical accuracy. I take personal responsibility for the quality of this document."

Signed: _____

Name: Russell Carter

Signed: _____

Name: Sam Fineberg

Branch Chief:

Approved: _____

Date & TR Number:

March 1994 RND-94-004